AD_____

Award Number: W81XWH-13-1-0028

TITLE:
Common Ground: An Interactive Visual Exploration and Discovery for Complex Health Data

PRINCIPAL INVESTIGATOR:
Yarden Livnat

CONTRACTING ORGANIZATION:
University of Utah
Salt Lake City, UT 84112

REPORT DATE:
April 2015

TYPE OF REPORT:
Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

x Approved for public release; distribution unlimited

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| April 2015 | Annual | 5 Mar 2014 – 4 Mar 2015 |

**4. TITLE AND SUBTITLE**
**Common Ground: An Interactive Visual Exploration and Discovery for Complex Health Data**

**5a. CONTRACT NUMBER**
W81XWH-13-1-0028

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Yarden Livnat, Per Gesteland, Adi Gundlapalli

E-Mail: yarden@sci.utah.edu, per.gesteland@hsc.utah.edu, Adi.Gundlapalli@hsc.utah.edu

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

UNIVERSITY OF UTAH, THE
201 S PRESIDENT CIRCLE RM 408
SALT LAKE CITY UT 84112-9023

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
The overarching objective of this work is to develop a novel, user-centric visual paradigm to enhance situational awareness by providing an effective visualization of large, complex and heterogeneous population health data. Presently, users of complex health data are overwhelmed with charts, graphs, tables and maps. Our goal is to develop a novel health data weather map visualization prototype that provides a dynamic, interactive presentation of complex healthcare data. We received a 6 months no cost extension and IRB extensions from both the University of Utah and Intermountain Healthcare. We received sample data from ESP and from Intermountain Healthcare. We have been working with Intermountain Healthcare on a new rich dataset extracted directly from medical notes using natural language processing (NLP) algorithms. Developed two visual analytics displays, the second of which is specifically designed for the new type of data.

**15. SUBJECT TERMS**
Visualization, Visual Analytics, Ontology, Software, Population Health Data

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | USAMRMC |
| U | U | U | UU | 16 | 19b. TELEPHONE NUMBER *(include area code)* |

**TABLE OF CONTENTS**

# 1 Introduction

The overarching objective of this work is to develop a novel, user-centric visual paradigm aimed at enhancing situational awareness by providing a clear, concise and effective visualization of large, complex and heterogeneous population health data. Our aim is to create a flexible and scalable population health visualization that depicts and distills the vast amount of data available from electronic health records using concise meta-data tags. Our goal is to further mature and evaluate an award winning prototype system we developed under the auspices of prior TATRC funding [1][2][3] We hypothesize that a well-designed visualization interface that is tailored to the users cognitive tasks, supports and promotes the discourse between users and their data and embodies domain knowledge will empower users to actively explore, enhance their ability to comprehend and analyze, and improve overall situational awareness. The project represents collaboration at the University of Utah between the Scientific Computing and Imaging Institute, the Department of Pediatrics, the Department of Medicine and the Department of Biomedical Informatics.

# 2 Body

We received on 02/11/15 a no-cost extension until September 4th 2015. We received an IRB extension approval from the University of Utah on May 15 2014 for the period ending on 5/15/2014. There was no gap in the IRB approval periods. We also received an IRB extension approval from Intermountain Healthcare for the period 07/31/2014 until 07/30/2015. There was a coverage gap in the IRB approval from Intermountain Healthcare as the original approval was until 5/10/2014. During the two months gap period we did use data from Intermountain Healthcare.

## 2.1 Data

### 2.1.1 ESP data

Both the design and the software development phases in this project rely on access to large and diverse health care data. In the proposal for this work we identified two sources for such data. The first source was the Early Stage Platform (ESP) for Medical Training and Health Information Sciences research and development that was developed by the Advanced Information Technology Group (AITG) in the Telemedicine and Advanced Technology Research Center (TATRC). The ESP data, described in the RFA for this grant, had great promise as it represented data that is closely aligned with the TATRC mission. The data was to be based on simulated population and thus it would have removed security and privacy concerns. In year one we were unable to gain access to the ESP data and we were informed by TATRC that the task of developing the ESP dataset would take much longer than anticipated.

Following our first yearly report, where we reported the issue with getting ESP data, we were referred to Ollie B. Gray a research program manager at HITG. We had a conference call on April 28$^{th}$ 2014 and we received small sample dataset containing information for 100 patients on the same day. We spend several weeks processing and analyzing the data. We recruited an undergrad student (Chaofeng Zhou) to write Python scripts to process the data and store it in an RDF repository. After further analysis of the data we came to the conclusion that most of the information is not applicable to our work. Some aspects of the data could in theory be used but after further analysis we concluded that the data resolution and specificity couldn't provide sufficient correlation and detail insights we need in for this project.

### 2.1.2 Intermountain Healthcare initial dataset

The second source for data that we identified in the proposal was Intermountain Healthcare. The advantage is that their data are based on healthcare operations data from a healthcare system that services a large segment of the population at the state of Utah. Gaining access to such data requires compliance with HIPPA regulations. The need for the date and zip code location data for each case patient for this project translates to the need for a limited data set (as opposed to a de-identified data set). In year one we worked with Intermountain Healthcare, the University of Utah IRB and TATRC IRB in order to get permission to use such data. The need for multiple agencies granting approval resulted in multiple delays but we received an initial dataset and developed a new graph database for it in August 2014.

### 2.1.3 Intermountain Healthcare dataset 2

During the second half of 2014, Intermountain Healthcare and Per Gesteland from our team have built a new database containing data for a set of 1,363,464 emergency department visits to six emergency departments in Salt Lake County spanning a seven-year period from 2007 to 2014. The data include basic demographic information (age, postal code, gender), clinical notes (Emergency department physician dictations), microbiological testing results (including testing for respiratory and enteric pathogens that commonly cause outbreaks) and diagnostic information (i.e., ICD-9 codes).

In November 2014, we received a subset of the data containing 1000 records (100 with confirmed influenza and 900 controls) that was used for free-text notes to natural language processing and probabilistic case detection (a Bayesian model for detecting cases of influenza). The team is currently working on processing the entire seven-year data set with the same technologies. This data present a unique and exciting opportunity to apply our novel visualization approach to rich and complex data and incorporate probability visualization into our display. The new dataset is different from previous data we used in that it contains both positive and *negative* findings as well as probabilities based on a state-of-the-art NLP classifiers. At that stage the data did not include geographic information or temporal information but we expect such data to be available as more data from the seven-year collection is processed. The new type of processed data is much better suited for the objectives of this project and includes exciting new dimensions such probabilities and negative findings. Now that a new software infrastructure at Intermountain Healthcare is up and running it is expected that we will receive more of the new type of data, including temporal and geospatial dimensions, on a regular and consistent bases.

### 2.1.4 Data repository

We evaluated various options for storing data such that we will be able to issue queries relating to both the health data and to the knowledgebase at the same time. In particular we looked at RDF repositories such as the open source Sesame (rdf4j.org), traditional relational databases and NoSQL type databases. We've used Sesame in previous projects to store ontologies and RDF based data but we concluded that Sesame might not scale well enough and that it doesn't offer the flexibility we need. We also looked at various types of NoSQL databases: key-value stores, column stores (e.g. Cassandra), document databases such as MongoDB and finally graph databases such as Neo4J.

We chose to implement or data repository using a graph database and in particular Neo4J database (http://www.neo4j.org). A graph database is a form of NoSQL database. NoSQL databases have gain popularity in big data and real-time web applications due to their simplicity, scalability and finer control. Since we did not have data from Intermountain Healthcare at that time, we used data we had from phase I of the project. We designed a graph-based representation for the data, converted the data,

meta-data and the knowledgebase to this format and finally deployed the new graph based repository on our server. We also created two additional graph databases for the ESP data and for the initial data from Intermountain Healthcare. Developed another graph database that encode and store initial data from Intermountain Health. The new data incorporate ICD-9 codes and we are working on pruning the data and converting ICD-9 codes to short text labels appropriate for our display.

Security is an important aspect and there are several issues in the design of Neo4J that we had to address. These issues relate only to the development stage and the development environment and would not be an issue in a production environment. The second dataset from Intermountain Healthcare, which we received in late November of 2014, is different in several aspects from previous data we obtained. The final format and scope of the available data has not been finalized. For this reason we opted to store this data in a standard relational database (MySQL).

## 2.2 Presentation

### 2.2.1 Software Prototype

The original software tool was developed as a small desktop application using Adobe Flex framework (http://www.adobe.com/products/flex.html). Our aim in this project is to develop a web-based application that can easily be access from any modern web browser. To achieve this goal our new design is based on a client-server architecture in which the server is responsible to all the communications with the data repository. The advantage of this approach is the decoupling between the client and the data repository. This in turn means that the client does not depend on a specific type or implementation details of the data repository. It also reduces security concerns as the data repository is kept behind a firewall and is accessible only by the dedicated server. The new architecture comprises of three layers: the data layer, middleware and a presentation layer.

### 2.2.2 User interface

In year one and part of year two we developed a browser based application based on server-client architecture. The code is written in JavaScript and HTML5 and was largely based on the AngularJS open source framework by Google (http://angularjs.org). AngularJS has gain large popularity for web development over the last few years.

#### *2.2.2.1 AngularJS*

As AngularJS gained more popularity more criticisms has started to be voiced about the steep learning curve, speed and complex architecture. These challenges proved to be major obstacles for us as well in part because there was no prior experience with it at the university research environment. During he development we found out that while there are various open source software components for AngularJS most were of low quality, unmaintained and very limited in their capabilities. For example, we worked to incorporate a map view based on the successful Leaflet (http://leafletjs.com) open software JavaScript library. The map view enables the user to see the spatial distribution of various subsets of the reported cases aggregated based on zip codes. The display of the zip codes layer poses a challenge when a user zooms in and out the map. To facilitate fast zooming capabilities we store several zip code shape files (http://en.wikipedia.org/wiki/Shapefile) at different resolution on our server and dynamically fetch the most appropriate one at run time. We successfully developed a proof of concept as a standalone tool but we encountered issues in integrating it into the AngularJS framework. We have used an open source bridging software that provide some support for using Leaflet based mapping in an AngularJS application but without much success. In late 2014 Google announced that version 2 of

AngularJS would not be back compatible with Angular 1.x and will represent a major architecture change.

Despite these challenges we were able to successfully develop an AngularJS based UI although the development cycle was much longer than anticipated. Nevertheless, for the second visualization software prototype we started to develop for the new Intermountain Healthcare data we received in November of 2014, we opt to do a complete architecture redesign. The new software does not use AngularJS anymore and instead use the D3 graphical library (see below) as the main framework. We used the D3 library in the first prototype but only for rendering the tag cloud and were very impressed with its capabilities.

#### *2.2.2.2 Tag cloud*

The main display features a dynamic graph view of tags in a form of a tag cloud and is built on top the D3 graphics library (http://d3js.org). The size of each tag represents the number of reported cases associated with that tag.

We improved our graph layout algorithm to better represent relationships between the visual items (tags). We developed a new correlation algorithm that facilitates time-based correlation between meta-data tags that are mutually exclusive (such as two cities or two age groups). In general we define an association between any pair of tags based on the number of shared items (reported cases) between them. The new algorithm adds new flexibility by allowing us to also define relationships between tags based on correlation relative to time.

We developed an algorithm to assist in preventing visual overlap between the visual items. The graph layout algorithm can employ different measures to determine the visual size of the tags and the correlation between them. We've developed new such measures that take into account the new probabilities information.

We are developing additional interactions and visual representations to encode both positive and *negative* findings. While the initial prototype only depicted information for positive findings, the new display enable a user to see the prevalence of negative findings and either focus on such patients or exclude them from the current selection.

### 2.3 Problems/Issues

The main problem we've been facing throughout the project is receiving data from Intermountain Healthcare. We have been working closely with them in the second half of 2014 but the process has been extremely slow. We received a small sample dataset in November of 2014 that on one hand includes exciting new features but it required us to develop a second visualization prototype. We were promised more data spanning seven years of data in early 2015 but it seems the delivery may be delayed once again. We are trying to work close with Intermountain Healthcare to ensure delivery as soon as possible.

## 3 Key Research Accomplishments

- Received sample data from ESP
- Received two different small datasets from Intermountain Healthcare

- Designed and deployed several graph databases to store the datasets
- Implemented a web-based prototype using AngularJS and D3
- Designed and implemented a second web-based prototype that does not rely on AngularJS and is designed to view and explore the new type of dataset we received from Intermountain Healthcare.
- Received IRB extension approvals from Intermountain Healthcare and University of Utah.
- Gave a presentation that included this work at Goldman Sachs and that was broadcasted live to their offices in SLC, Austin Texas, NYC, London and Bangalore.
- Received an NSF SBIR grant ($750K/2 years). The work is based in part of the exploration display approach we developed under this grant.

## 4 Reportable Outcome

### 4.1 Publications

### 4.2 Presentations

Y. Livnat, Visual Exploration for Situational Awareness

- Goldman Sachs, 2015

### 4.3 Informatics

Developed graph databases (using Neo4J) for sample data from ESP and for the first dataset we received from Intermountain Healthcare. Created a relational database (using MySQL) for the second dataset from Intermountain Healthcare.

### 4.4 Funding based on this work

The following are NSF funding based on the visualization we developed as part of work. The proposals demonstrate the generality of this visual paradigm to wide range of other domains.

1. NSF SBIR Phase II
   PI: Brad Davis (Enclavix), Yarden Livnat (UofU PI)
   Budget: $750K/2 years (UofU: $375K/2 years)
   Title: Automated System to Identify and Curate Web-based Resources for Entrepreneurs
   Status: Funded. Work began in October 2014.

## 5 Conclusion

We have made significant progress on the objectives of this project. With the goal of maturing a prototype for enhancing situational awareness by effective visualization of large, complex and heterogeneous population health data, we have in year two (1) received several small datasets and developed data repositories to represent them (2) made progress in obtaining healthcare data sets from a large operational partner (Intermountain Healthcare); (3) developed two software prototypes including one that represent a complete architecture of the software.

While the specific domain of interest to this project is bio-surveillance with a focus on respiratory infections, we note that the scientific principles of user-centric design and contextual inquiries are portable to other clinical domains. Our results and deliverables based on applying sound informatics

techniques and principles to visualize and explore large, complex and heterogeneous population health data will serve as viable models for analysis of big healthcare data

## 6 Reference


[1] A. V. Gundlapalli, Y. Livnat, and P. H. Gesteland, "Final Report Submitted to U.S. Army Medical Research and Materiel Command: Visual Correlation for the Early Detection of Infectious Disease Outbreaks Award Number W81XWH0710699)," University of Utah School of Medicine, Salt Lake City, UT, 2009.

[2] P. Gesteland, Y. Livnat, N. Galli, M. H. Samore, A. V. Gundlapalli. "The EpiCanvas Infectious Disease Weather map: An Interactive Visual Exploration of Temporal Correlations". Journal of the American Medical Informatics Association (JAMIA), Vol. 9, pp. 954-959, 2012. [**ISDS Award for Outstanding Research Article in Biosurveillance (Scientific Achievement)** selected from all journal publications since 2010, ISDS 2012

[3] P. Gesteland, Y. Livnat, N. Galli, M. H. Samore, A. V. Gundlapalli. "The EpiCanvas Infectious Disease Weather map: An Interactive Visual Exploration of Temporal Correlations". AMIA 2011 Annual Symposium, 2011 **(Winner of the Homer R. Warner Award)**


## 7 Appendices

None